Explore and Summarise Data

Veena Reddy Hanumanthgari

April 8, 2018

This report explores a dataset containing loan data and borrower details for approximately 114,000 listings

```
chooseCRANmirror(graphics = FALSE, ind = 1)
knitr::opts_chunk$set(fig.path = 'Figs/', echo = FALSE, message = FALSE, warning
                      = FALSE)
```

Univariate Plots Section:

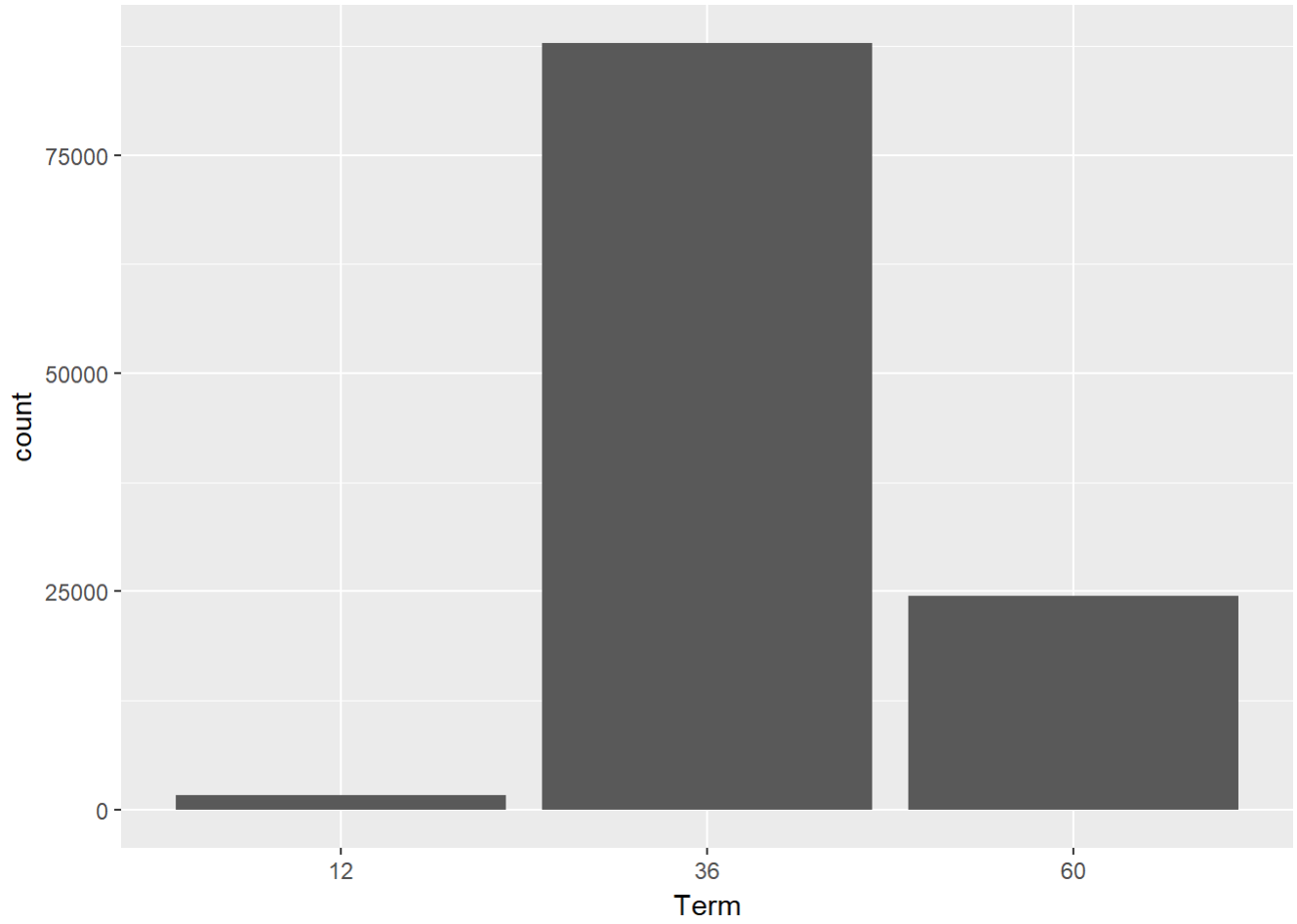
```
## [1] 113937     12
```

```
## 'data.frame':    113937 obs. of  12 variables:
##  $ ListingNumber      : int  544844 630052 576640 530423 483095 530332 647225 562388 944577 4
52658 ...
##  $ Term               : int  36 36 36 36 36 36 36 36 36 36 ...
##  $ LoanStatus         : Factor w/ 12 levels "Cancelled","Chargedoff",..: 5 3 4 4 3 3 3 4 4 3
...
##  $ ProsperScore       : int  1 1 1 1 1 1 1 1 1 ...
##  $ ListingCategory    : int  2 18 18 1 1 3 19 20 1 3 ...
##  $ BorrowerState      : Factor w/ 52 levels "","AK","AL","AR",..: 25 6 33 6 6 26 45 47 16 11
...
##  $ LoanOriginalAmount : int  5500 4000 2500 4000 1500 4000 3000 3000 15000 2000 ...
##  $ LoanOriginationDate: Factor w/ 1873 levels "1/10/2006","1/10/2007",..: 520 1670 973 299 38
4 295 213 718 587 988 ...
##  $ Investors          : int  45 45 12 52 27 88 17 1 1 60 ...
##  $ StatedMonthlyIncome: num  3083 6125 2083 2875 9583 ...
##  $ IncomeVerifiable   : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ IncomeRange        : Factor w/ 8 levels "$0 ","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4 ...
```

```
##  ListingNumber          Term                       LoanStatus
##  Min.   :      4   Min.   :12.00   Current              :56576
##  1st Qu.: 400919   1st Qu.:36.00   Completed            :38074
##  Median : 600554   Median :36.00   Chargedoff           :11992
##  Mean   : 627886   Mean   :40.83   Defaulted            : 5018
##  3rd Qu.: 892634   3rd Qu.:36.00   Past Due (1-15 days) :  806
##  Max.   :1255725   Max.   :60.00   Past Due (31-60 days):  363
##                                    (Other)              : 1108
##   ProsperScore   ListingCategory  BorrowerState   LoanOriginalAmount
##  Min.   : 1.00   Min.   : 0.000   CA     :14717   Min.   : 1000
##  1st Qu.: 4.00   1st Qu.: 1.000   TX     : 6842   1st Qu.: 4000
##  Median : 6.00   Median : 1.000   NY     : 6729   Median : 6500
##  Mean   : 5.95   Mean   : 2.774   FL     : 6720   Mean   : 8337
##  3rd Qu.: 8.00   3rd Qu.: 3.000   IL     : 5921   3rd Qu.:12000
##  Max.   :11.00   Max.   :20.000          : 5515   Max.   :35000
##  NA's   :29084                    (Other):67493
##  LoanOriginationDate   Investors       StatedMonthlyIncome
##  1/22/2014 :   491   Min.   :   1.00   Min.   :      0
##  11/13/2013:   490   1st Qu.:   2.00   1st Qu.:   3200
##  2/19/2014 :   439   Median :  44.00   Median :   4667
##  10/16/2013:   434   Mean   :  80.48   Mean   :   5608
##  1/28/2014 :   339   3rd Qu.: 115.00   3rd Qu.:   6825
##  9/24/2013 :   316   Max.   :1189.00   Max.   :1750003
##  (Other)   :111428
##  IncomeVerifiable         IncomeRange
##  Mode :logical    $25,000-49,999:32192
##  FALSE:8669       $50,000-74,999:31050
##  TRUE :105268     $100,000+     :17337
##                   $75,000-99,999:16916
##                   Not displayed : 7741
##                   $1-24,999     : 7274
##                   (Other)       : 1427
```

The dataset consists of 10 variables, with almost 114,000 observations.

TERM:




```
##
##    12    36    60
##  1614 87778 24545
```

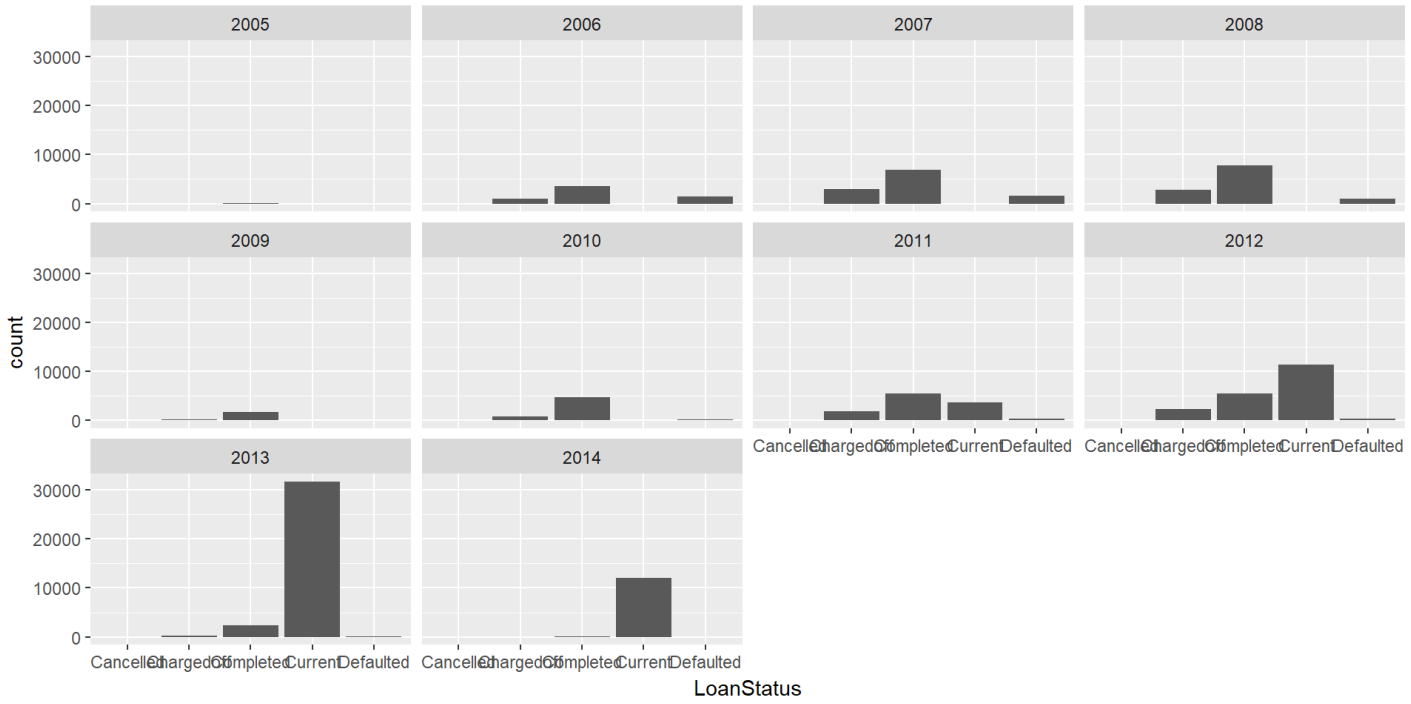
Creating another column for year of loan origination:

```
##   ListingNumber Term LoanStatus ProsperScore ListingCategory BorrowerState
## 1        544844   36  Defaulted            1               2            MN
## 2        630052   36  Completed            1              18            CA
## 3        576640   36    Current            1              18            NJ
## 4        530423   36    Current            1               1            CA
## 5        483095   36  Completed            1               1            CA
## 6        530332   36  Completed            1               3            MO
##   LoanOriginalAmount LoanOriginationDate Investors StatedMonthlyIncome
## 1               5500          12/20/2011        45            3083.333
## 2               4000           8/30/2012        45            6125.000
## 3               2500           4/18/2012        12            2083.333
## 4               4000           10/7/2011        52            2875.000
## 5               1500          11/23/2010        27            9583.333
## 6               4000           10/6/2011        88            8333.333
##   IncomeVerifiable    IncomeRange LoanOriginationYear
## 1             TRUE $25,000-49,999                2011
## 2             TRUE $50,000-74,999                2012
## 3             TRUE  Not displayed                2012
## 4             TRUE $25,000-49,999                2011
## 5             TRUE      $100,000+                2010
## 6             TRUE      $100,000+                2011
```

LOAN STATUS:

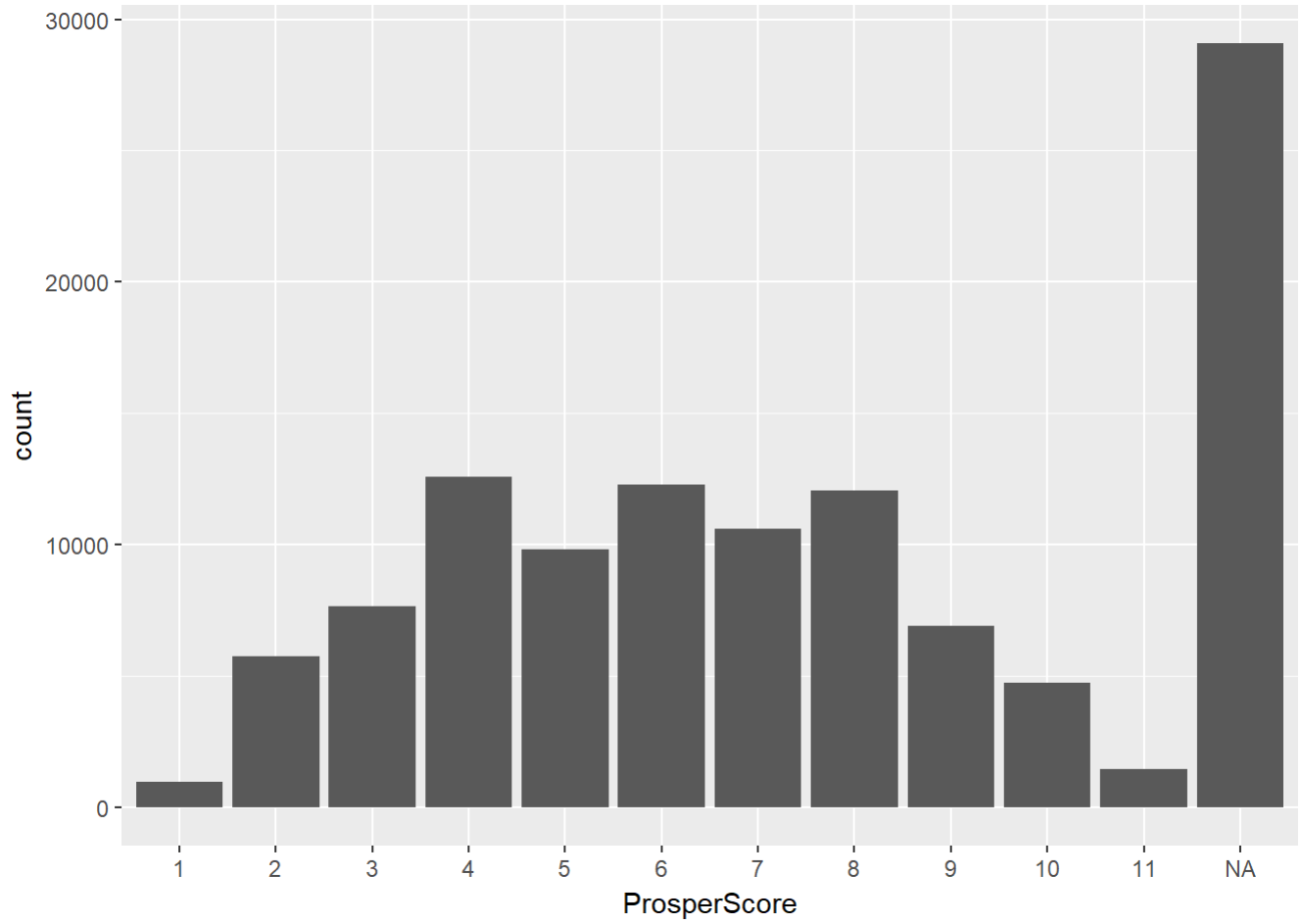
```
##
##              Cancelled            Chargedoff              Completed
##                      5                 11992                  38074
##                Current              Defaulted FinalPaymentInProgress
##                  56576                   5018                    205
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days)
##                     16                    806                    265
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days)
##                    363                    313                    304
```

```
##  Cancelled Chargedoff  Completed    Current  Defaulted
##          5      11992      38074      58848       5018
```



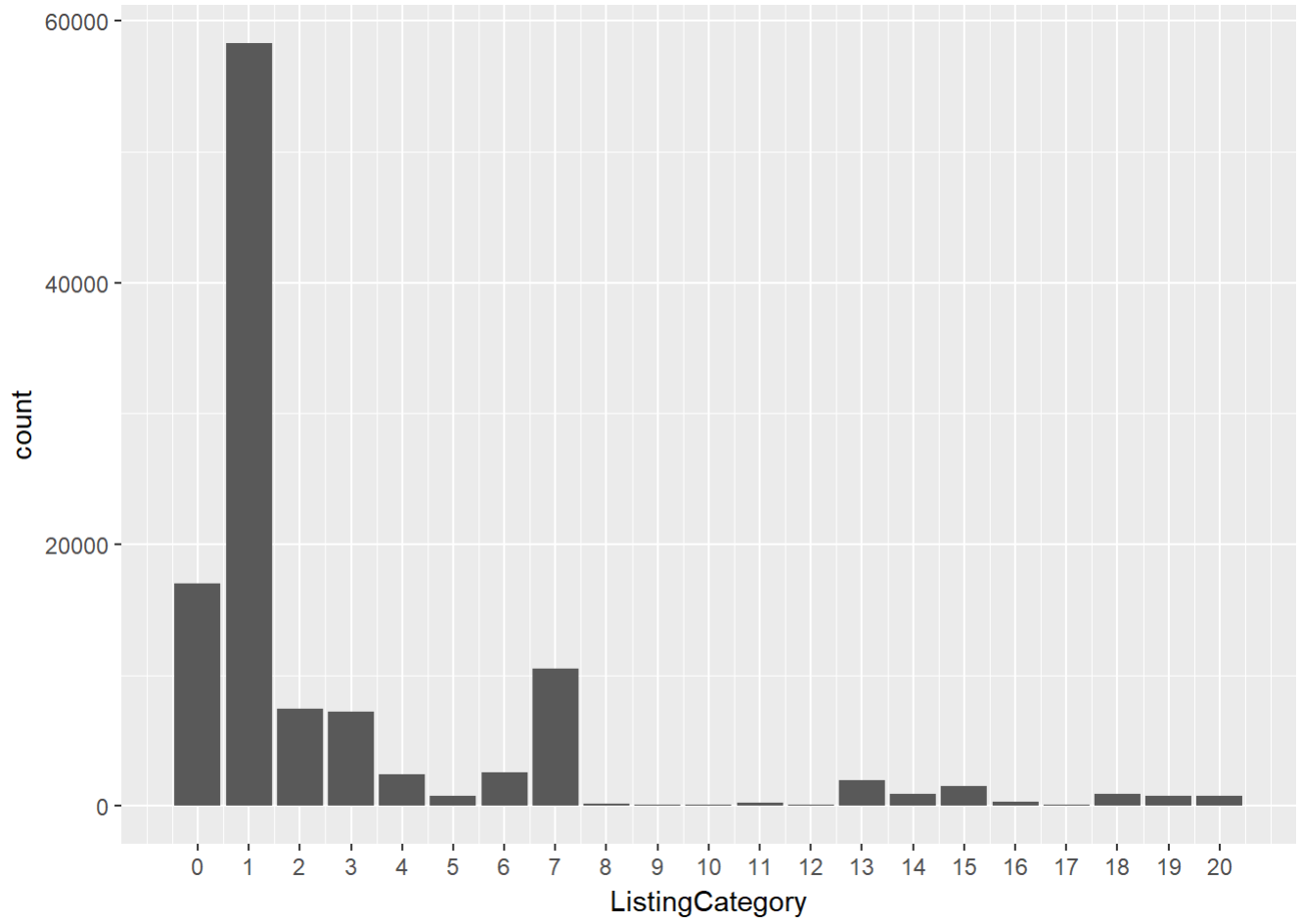
For ease of analysis, I have changed the loan status of all listings ‘past due’ to ‘current’. By splitting the data over LoanOrginationYear using facet_wrap, we get a clear picture of counts of loan status of the data for each year. For the year 2005, there are only 22 records and hence do not show up on the histogram. And for the year 2014, data is available only for Jan, Feb and March.

PROSPER SCORE:A custom risk score built using historical Prosper data. The score ranges from 1-10, with 10 being the best, or lowest risk score. Applicable for loans originated after July 2009.




```
##
##    1    2    3    4    5    6    7    8    9   10   11
##  992 5766 7642 12595 9813 12278 10597 12053 6911 4750 1456
```

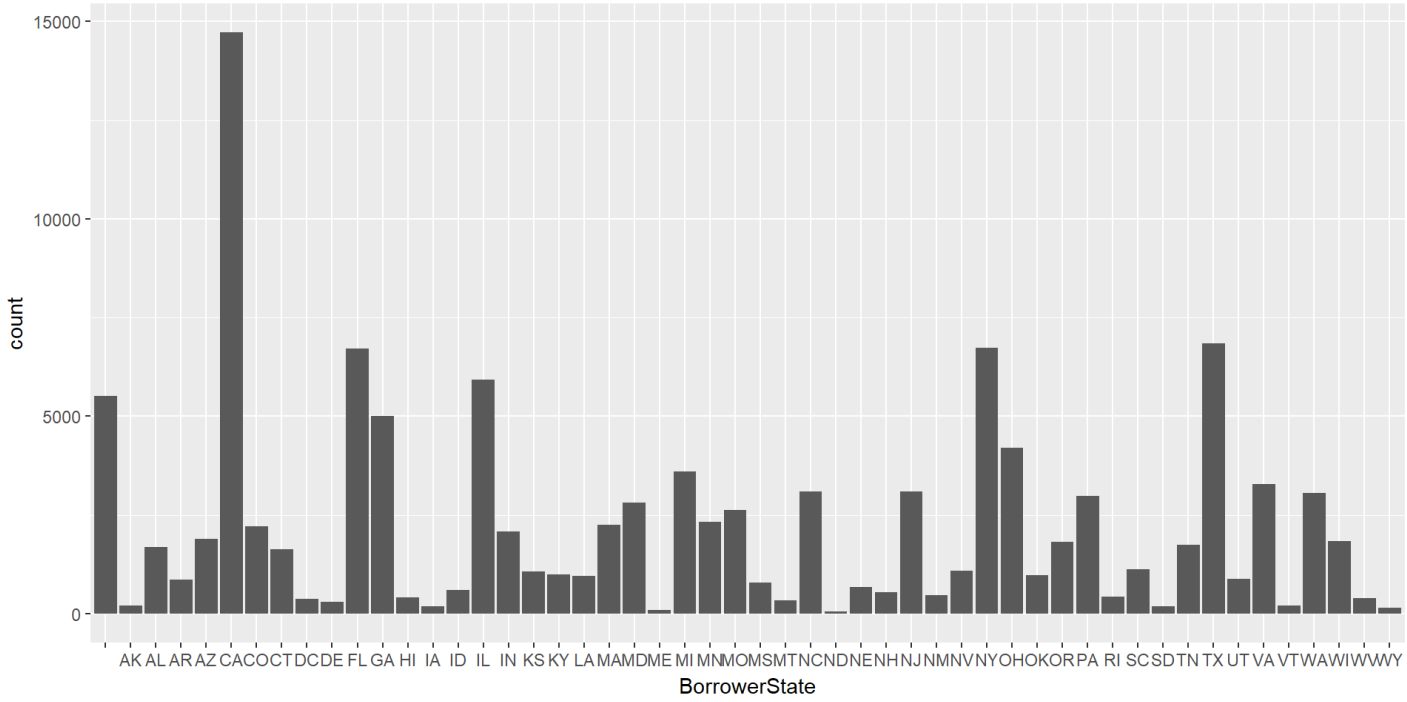
LISTING CATEGORY: The category of the listing that the borrower selected when posting their listing:




```
##
##     0     1     2     3     4     5     6     7     8     9    10    11
## 16965 58308  7433  7189  2395   756  2572 10494   199    85    91   217
##    12    13    14    15    16    17    18    19    20
##    59  1996   876  1522   304    52   885   768   771
```

Debt Consolidation is the category for which maximum number of listings were created.Considerable number of listings have NA or other as their Category. Home Improvement and business are the categories mentioned for the next highest number of listings.

BORROWER STATE:

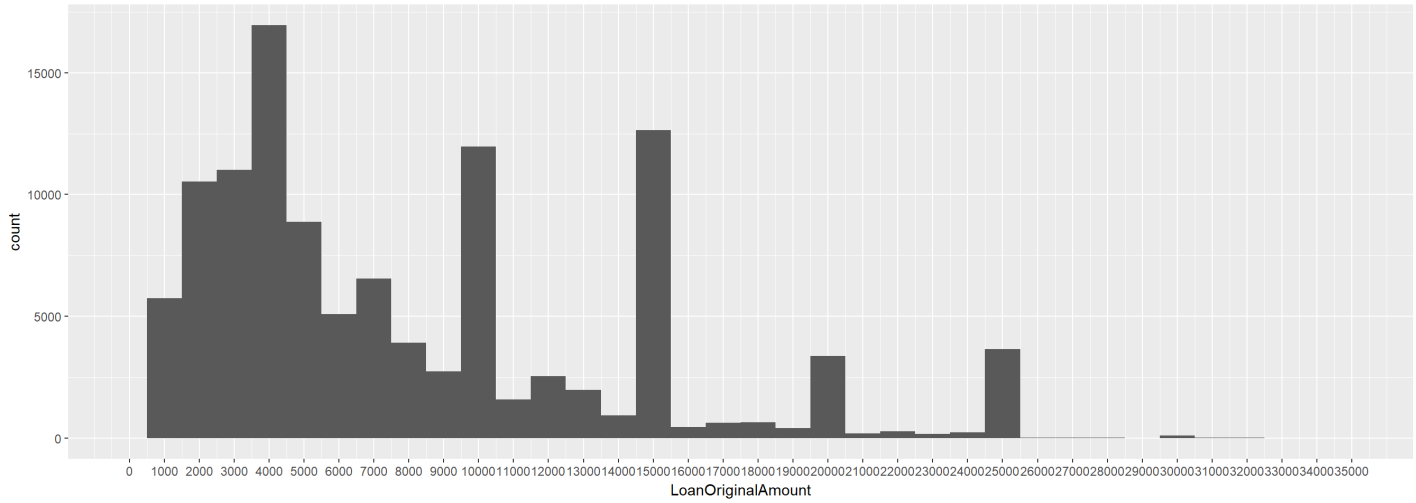



```
##
##          AK    AL    AR    AZ    CA    CO    CT    DC    DE    FL    GA
##  5515   200  1679   855  1901 14717  2210  1627   382   300  6720  5008
##    HI    IA    ID    IL    IN    KS    KY    LA    MA    MD    ME    MI
##   409   186   599  5921  2078  1062   983   954  2242  2821   101  3593
##    MN    MO    MS    MT    NC    ND    NE    NH    NJ    NM    NV    NY
##  2318  2615   787   330  3084    52   674   551  3097   472  1090  6729
##    OH    OK    OR    PA    RI    SC    SD    TN    TX    UT    VA    VT
##  4197   971  1817  2972   435  1122   189  1737  6842   877  3278   207
##    WA    WI    WV    WY
##  3048  1842   391   150
```

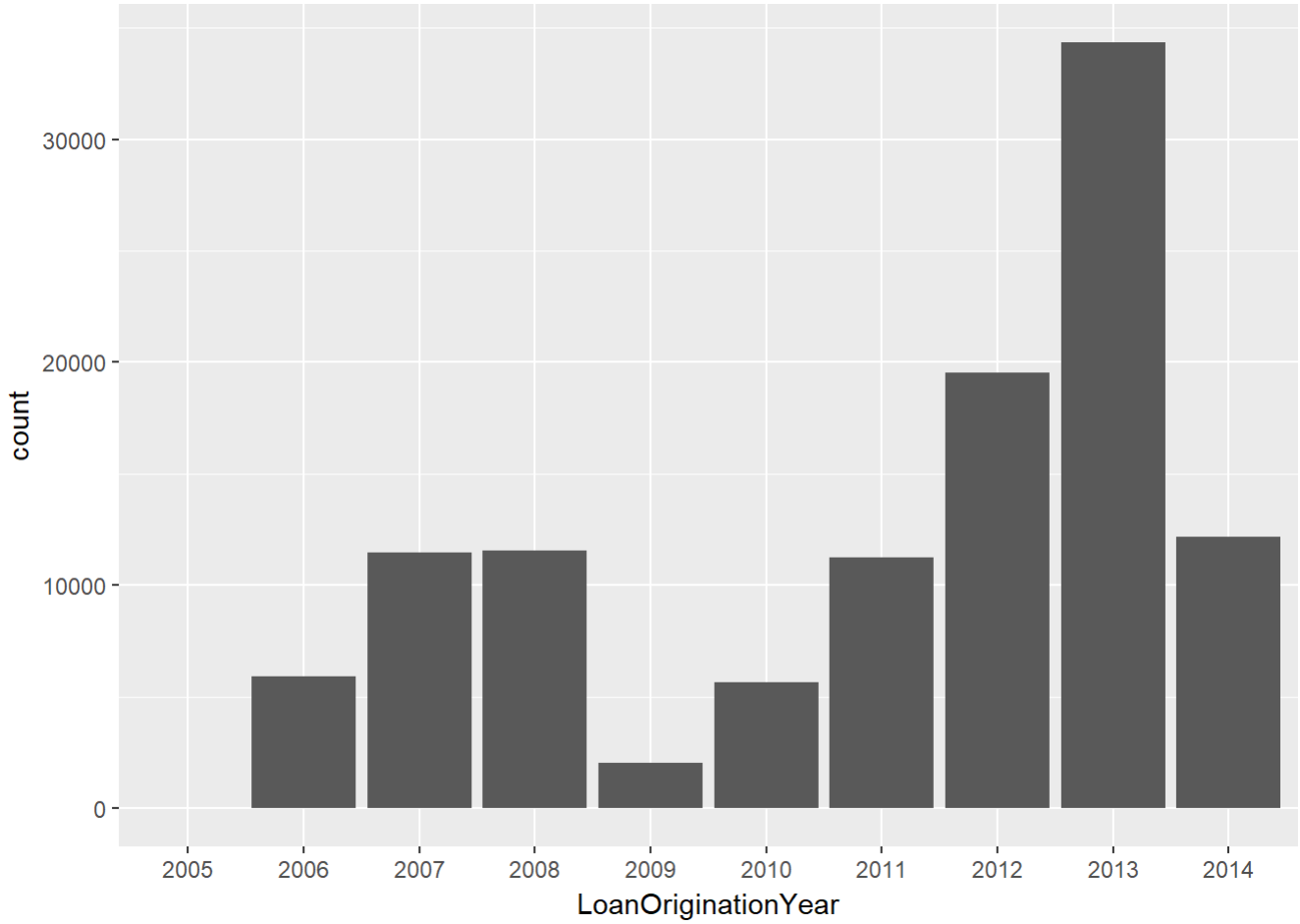
As can be seen form the plot,the state wise count for listings is highest for the home state California with Florida, New York and Texas lining up next. Illinois, Georgia and Ohio also have a good count.

LOAN ORIGINAL AMOUNT:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    6500    8337   12000   35000
```



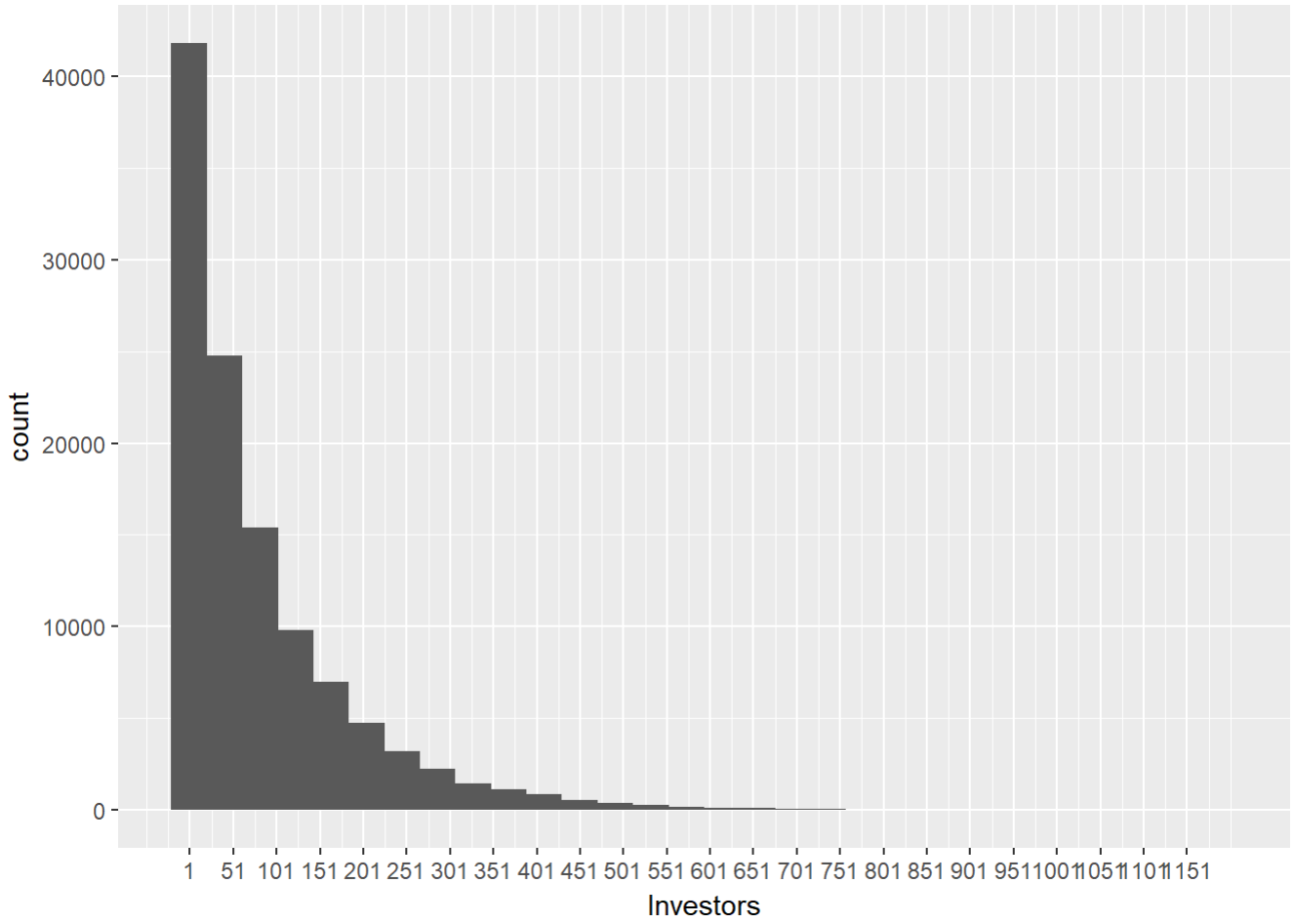
LOANORIGINATIONYEAR:



There is a clear pattern in the year-wise barplot for the number of loans borrowed each year. Again there a a very few loans for the year 2005, the number has increased for two consecutive years, remained constant for 2008, an then there is a sudden drop in the year 2009. From there on, the count has exponentially been increasing. Note that we have only 3 months data for the year 2014.

INVESTORS:

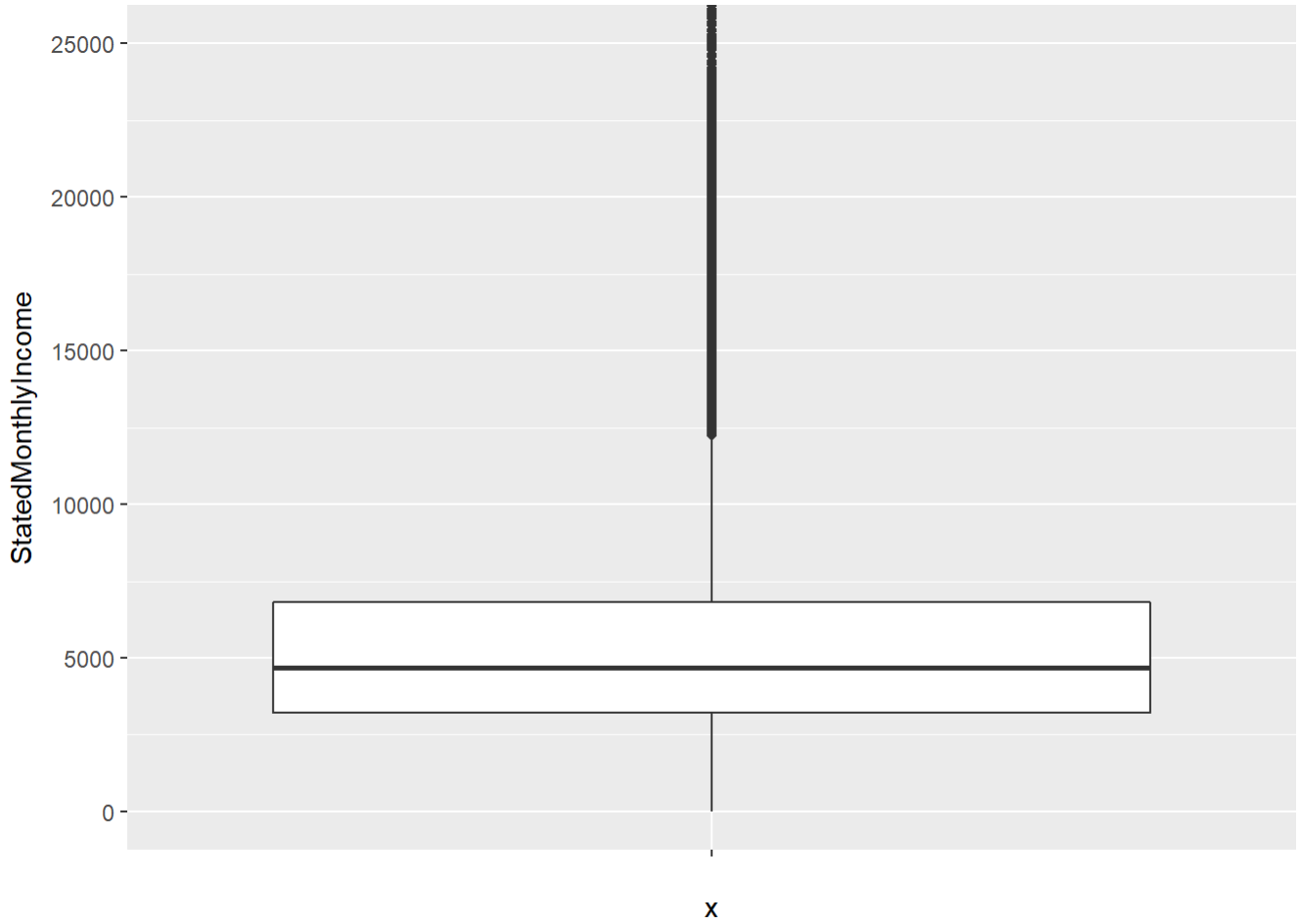
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00   44.00   80.48  115.00 1189.00
```



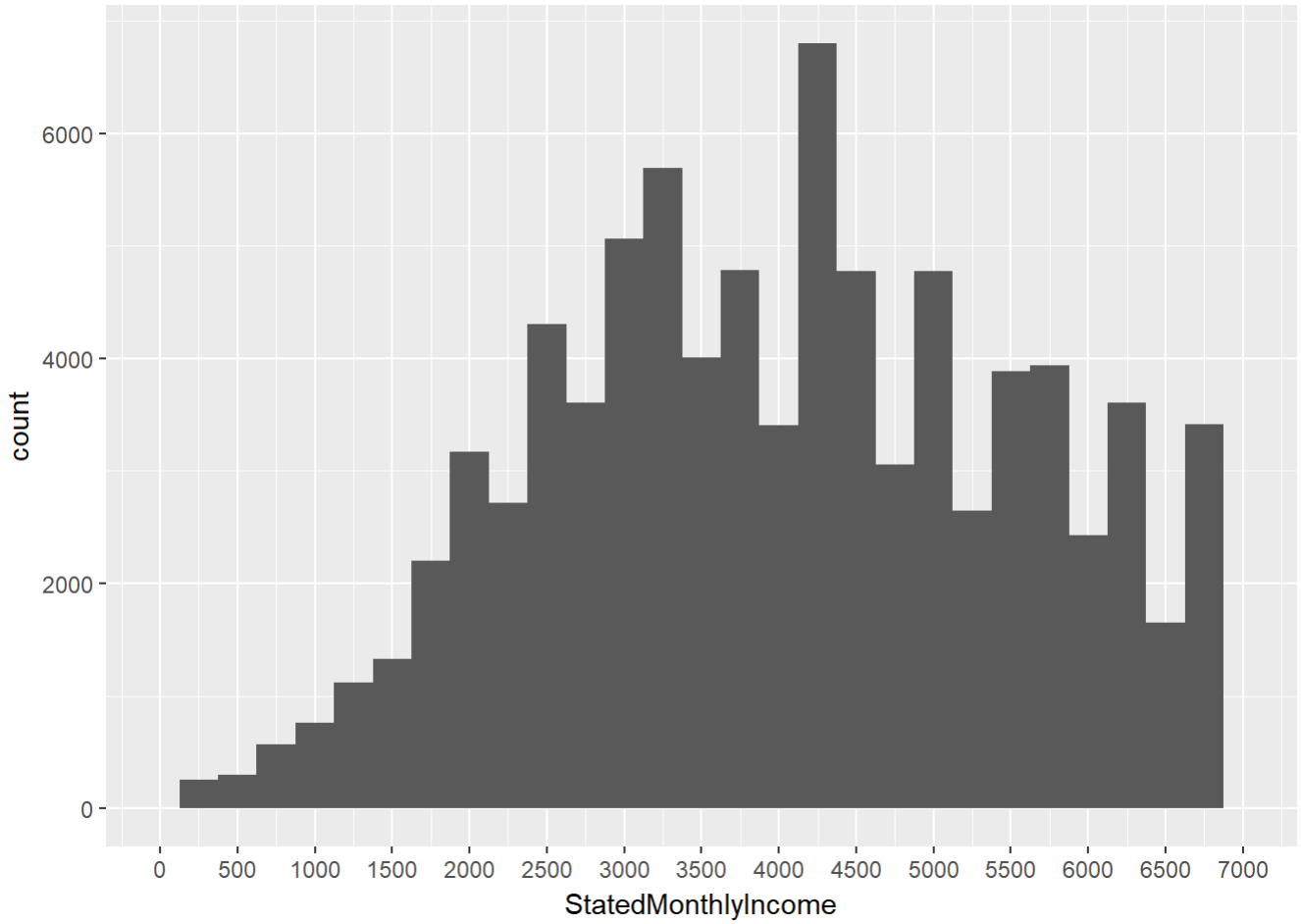
This data is long-tailed and skewed to the right. The median value for number of investors is 44.

STATED MONTHLY INCOME:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3200    4667    5608    6825 1750003
```



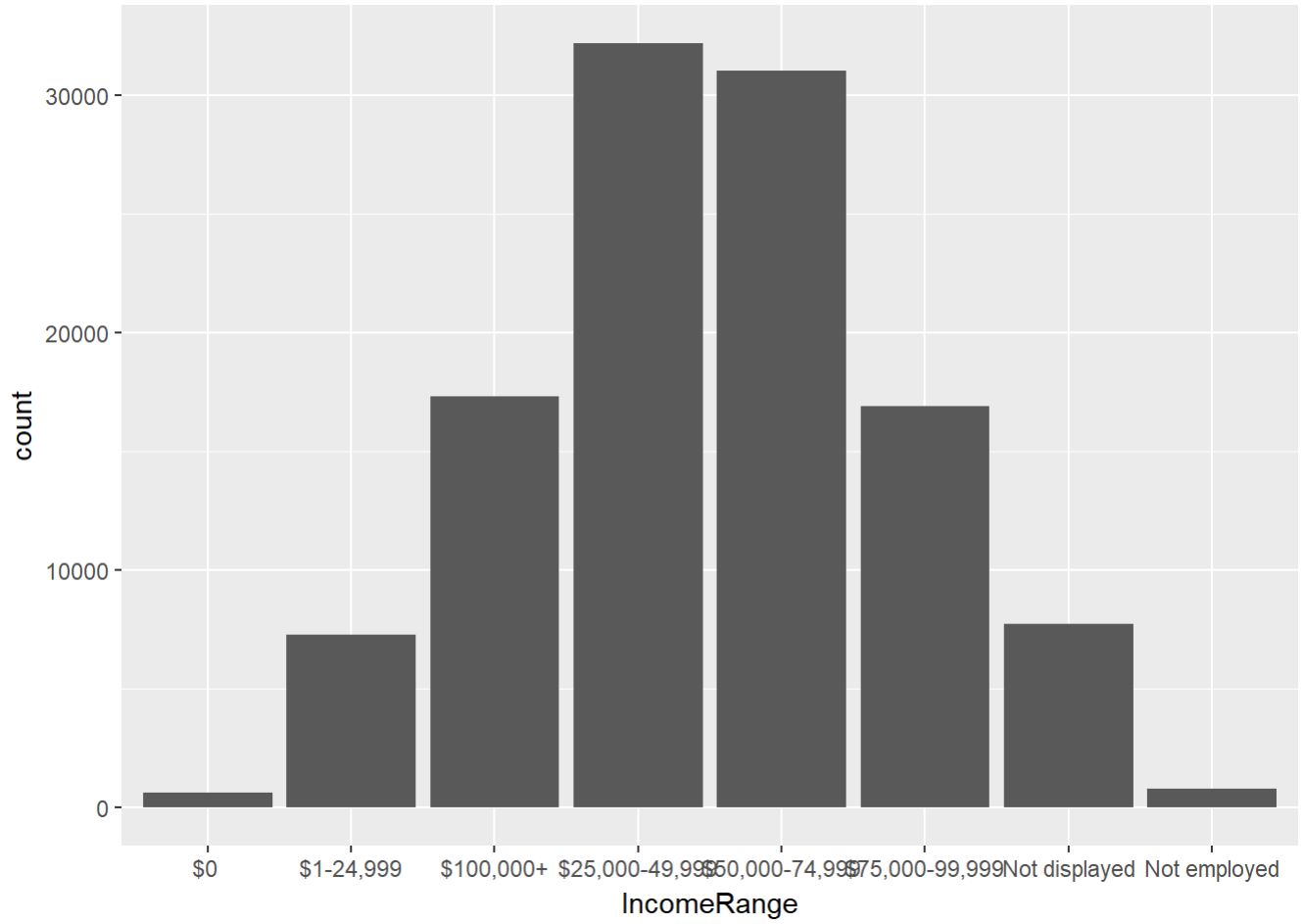
Looking at the summary/boxplot,it is obvious that this data is very dispersed with considerable amount of the data in the outlier section. To get a better picture of income counts under the third quartile, we plot a histogram below.



INCOME RANGE:

```
##            $0       $1-24,999      $100,000+ $25,000-49,999 $50,000-74,999
##           621            7274          17337          32192          31050
## $75,000-99,999  Not displayed   Not employed
##          16916            7741            806
```

```
##
##            $0       $1-24,999      $100,000+ $25,000-49,999 $50,000-74,999
##           621            7274          17337          32192          31050
## $75,000-99,999  Not displayed   Not employed
##          16916            7741            806
```



Univariate Analysis:

Structure of Dataset:

There are 113,937 listings in the dataset with 10 features(ListingNumber, Term, LoanStatus, ProsperScore, ListingCateogry, BorrowerState, LoanOriginalAmount, LoanOriginationDate, Investors, StatedMonthlyIncome, IncomeVerifiable, IncomeRange). The variables Term, LoanStatus, ProsperScore, ListingCategory, BorrowerState, IncomeVerifiable and IncomeRange are factor variables with the following levels.

Term: 12,36,60(in months)

LoanStatus:Cancelled,Chargedoff,Completed,Current,Defaulted

ProsperScore:1-10, with 10 being the best, or lowest risk score. ListingCategory: 0 - Not Available 1 - Debt Consolidation, 2 - Home Improvement, 3 - Business, 4 - Personal Loan, 5 - Student Use, 6 - Auto, 7- Other, 8 - Baby&Adoption, 9 - Boat, 10 - Cosmetic Procedure, 11 - Engagement Ring, 12 - Green Loans, 13 - Household Expenses, 14 - Large Purchases, 15 - Medical/Dental, 16 - Motorcycle, 17 - RV, 18 - Taxes, 19 - Vacation, 20 - Wedding Loans

BorrowerState: All US States.

Income Verifiable:True,False

Income Range:$0, $1-24,999, $25,000-49,999, $50,000-74,999, $75,000-99,999, $100,000+, Notdisplayed, Not employed.

Other observations:

-Most listings have a term of 3 years.

-Most common Listing Category mentioned is Debt Consolidation.

-The median value of Loan Original Amount is $6500.

-The median value of StatedMonthlyIncome value is $4667.

-California is the state with highest number of borrowers given any year.

Main features of interest in the dataset:

The main features in the dataset are LoanOriginalAmount, BorrowerState and LoanOriginationYear.I would like to observe and analyse the variation of loans for different states and over the years.

Created 1 new variable from existing variables in the dataset:

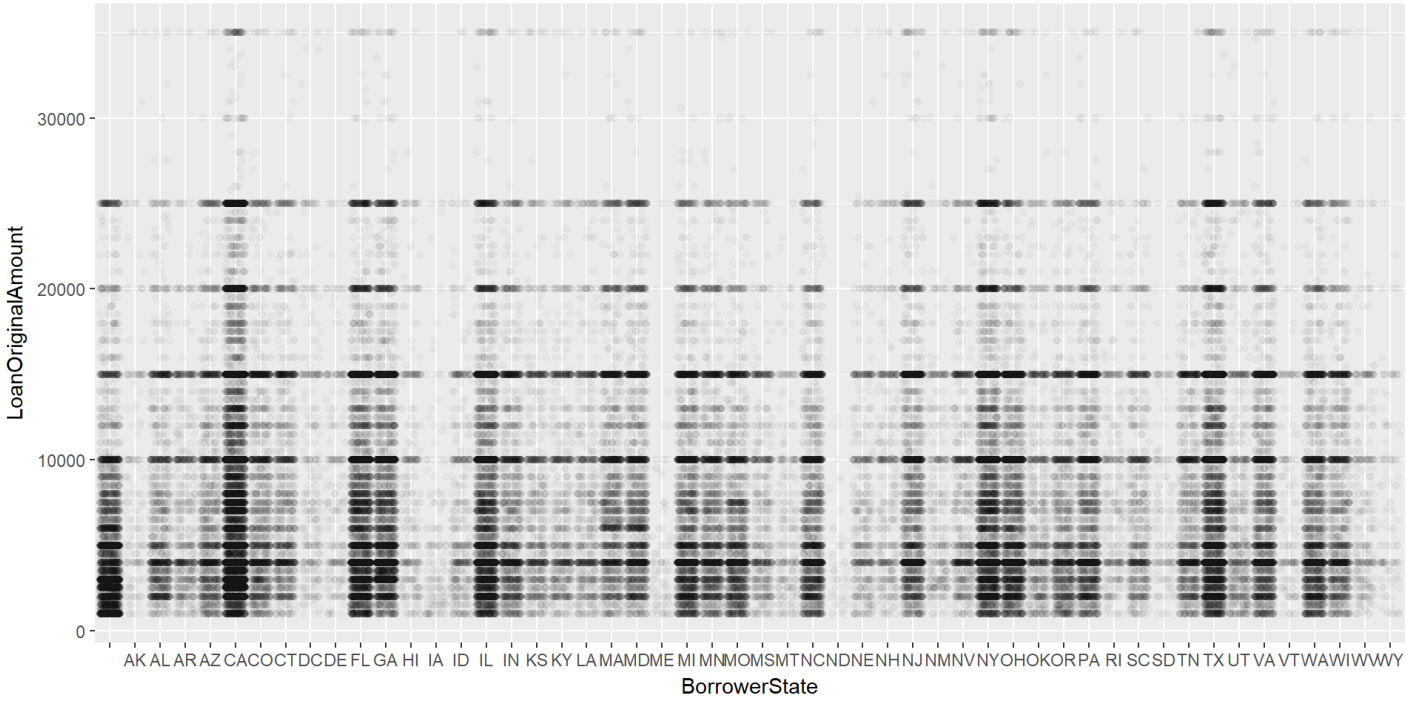
I created 1 new variable ‘LOanOrginationYear’ by extracting the year from ‘LoanOriginationDate’ variable. This is to analyse the variation of some of the other variables over the years.

Unusual distributions:

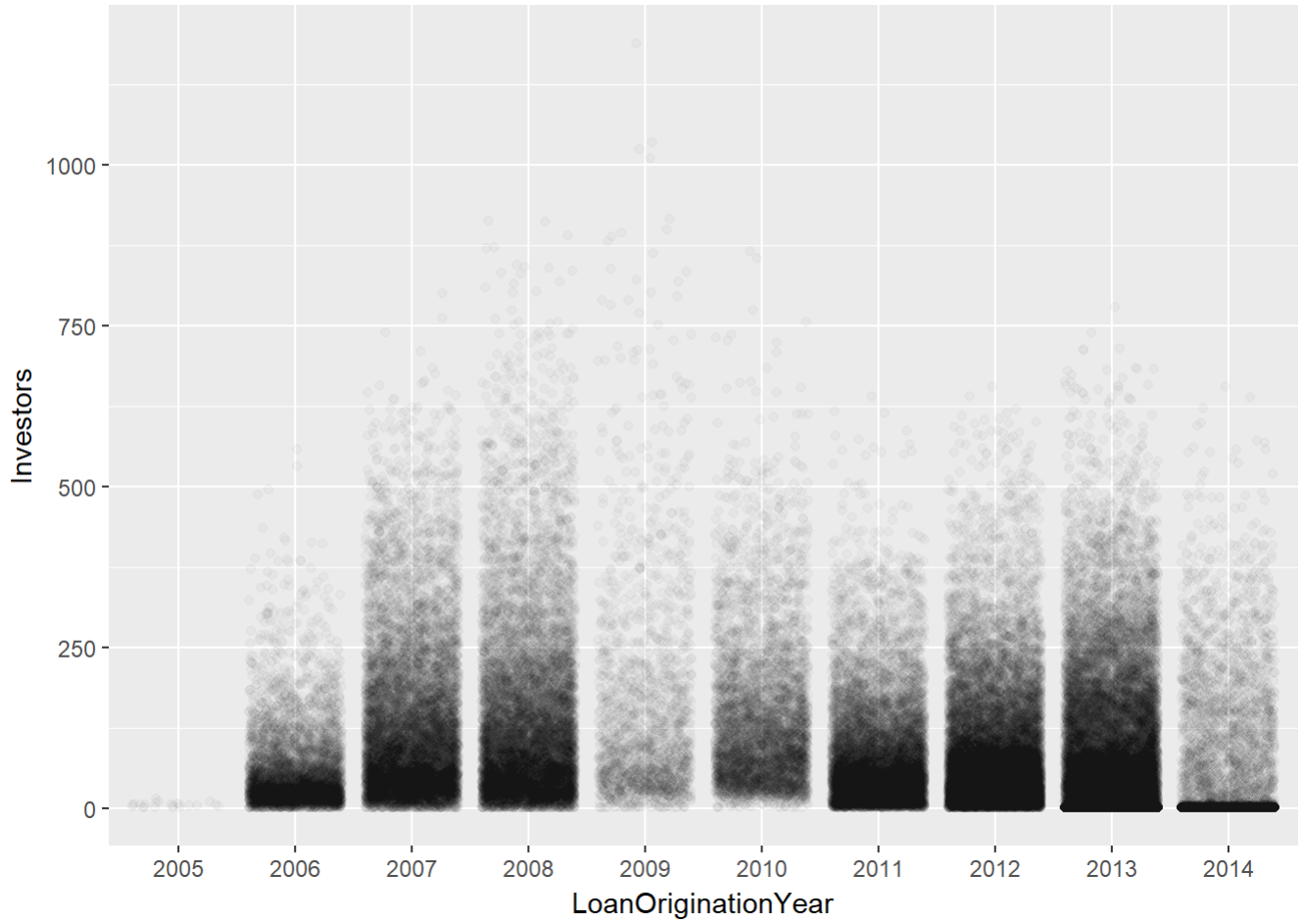
I have also changed some values of factor variable ‘LoanStatus’. All the listings that are past due date but not yet defaulted have been changed to ‘Current’.This has been performed just to tidy the data and aid with graphical analysis of Loan Status vs other variables of the listings. There are listings with very high Stated Monthly Incomes ranging from 100K-600K.

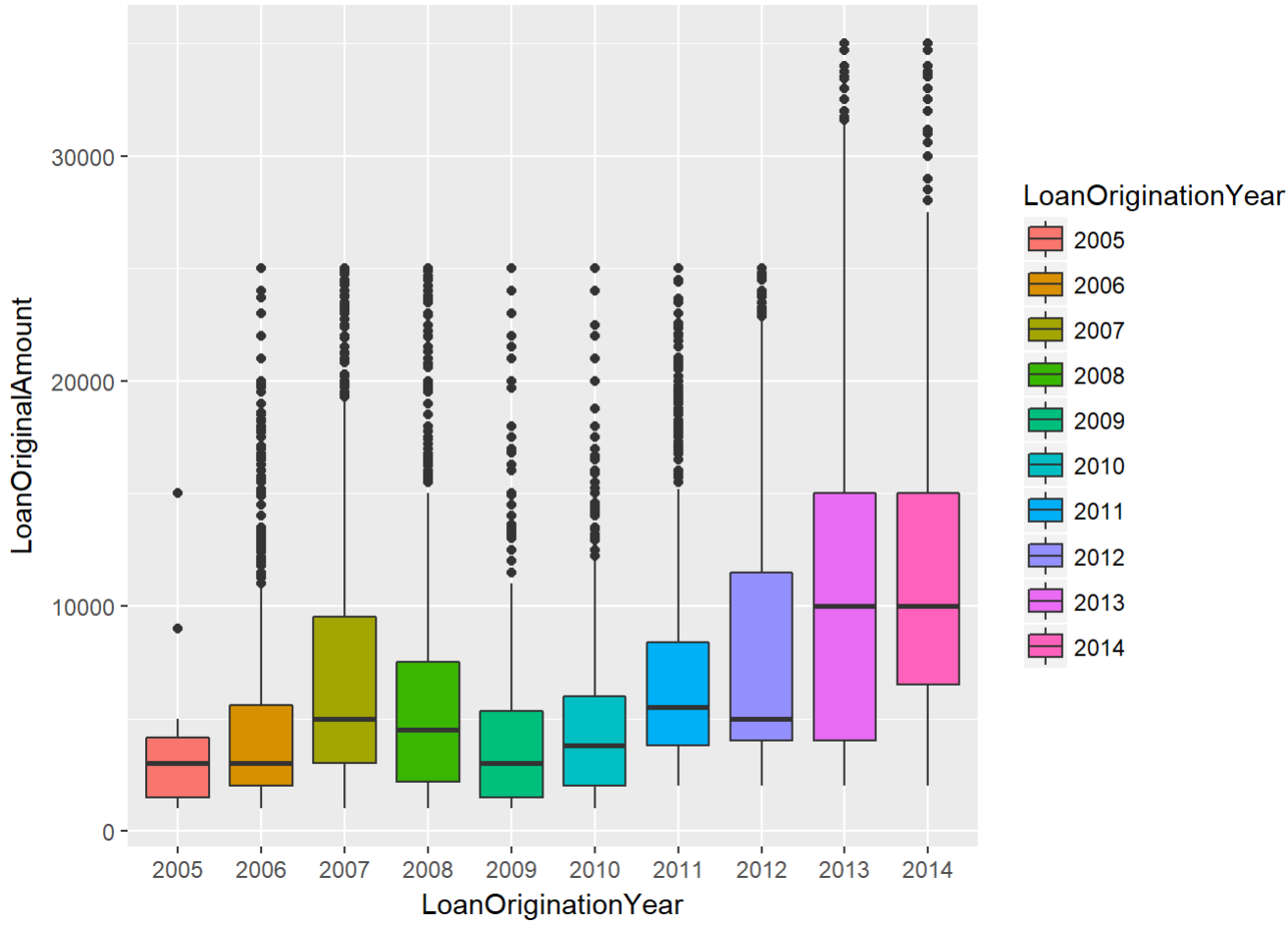
Bivariate Plots Section

For years 2005 and 2006, Listing Category is not available.And from 2007 to 2010, out of the 7 available categories, most listings in each year have been created for debt consolidation.Home Improvement and Business categories also have a good proportion of listings. The count of listings with larger loan amounts have also increased for the years 2013 and 2014. Higher loan amounts can also be seen for other categories in these two years. This shows increased interest of borrowers as well as investors in Prosper.

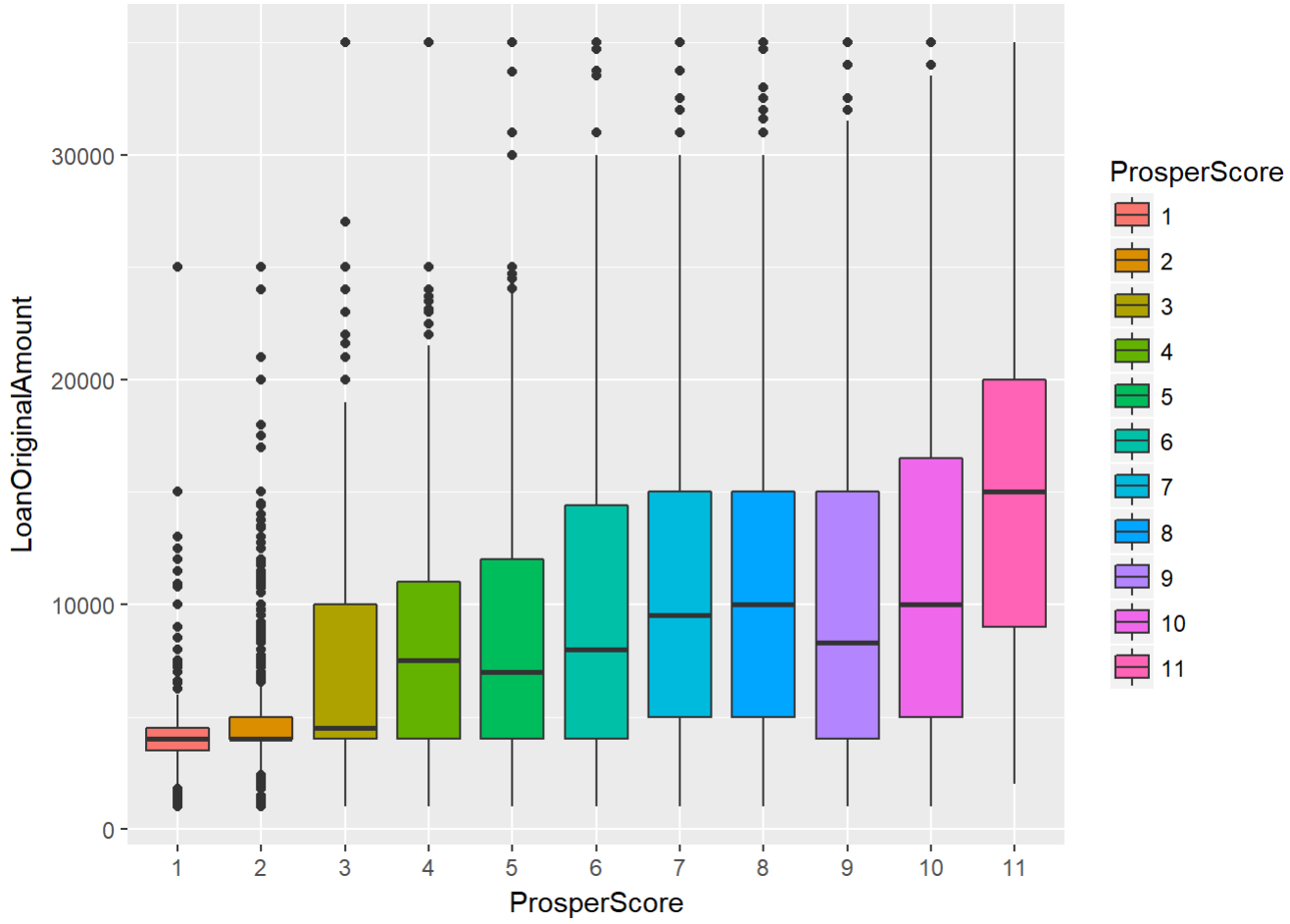


Most loan amounts, even for the states with highest number of borrowers, are below $10,000. And clear horizontal lines can be seen at the regular numbers like $10K, $15K, $20K and $25K but the dots get lighter going up the y-axis.


Investors
1000
750
500
250
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
LoanOriginationYear


LoanOriginalAmount
30000
20000
10000
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
LoanOriginationYear
LoanOriginationYear
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

The same trend as the one above for Investors can be seen here for median values except for the year 2012 where there is an unusual drop in median Loan Original Amount.



Bivariate analysis

Debt Consolidation is the topmost mentioned Listing Category, followed by Home Improvement and Business. Household Expenses, Large Purchases, Medical/Dental, Motorcycle,Taxes, Vacation and Wedding Loans were also mentioned in other listings but are less popular and have loan amounts $15000 and under.

The drop in median value of LoanOriginalAmount in the year 2012 for which the cause is unknown.There is also a sudden increase in higher loan amounts visible in the outliers area of the plot in the year 2013 and continues in 2014 also.

States with most number of borrowers are CA, NY ,TX and FL. States like AK,ME,ND,WY and IA have the least number of borrowers.

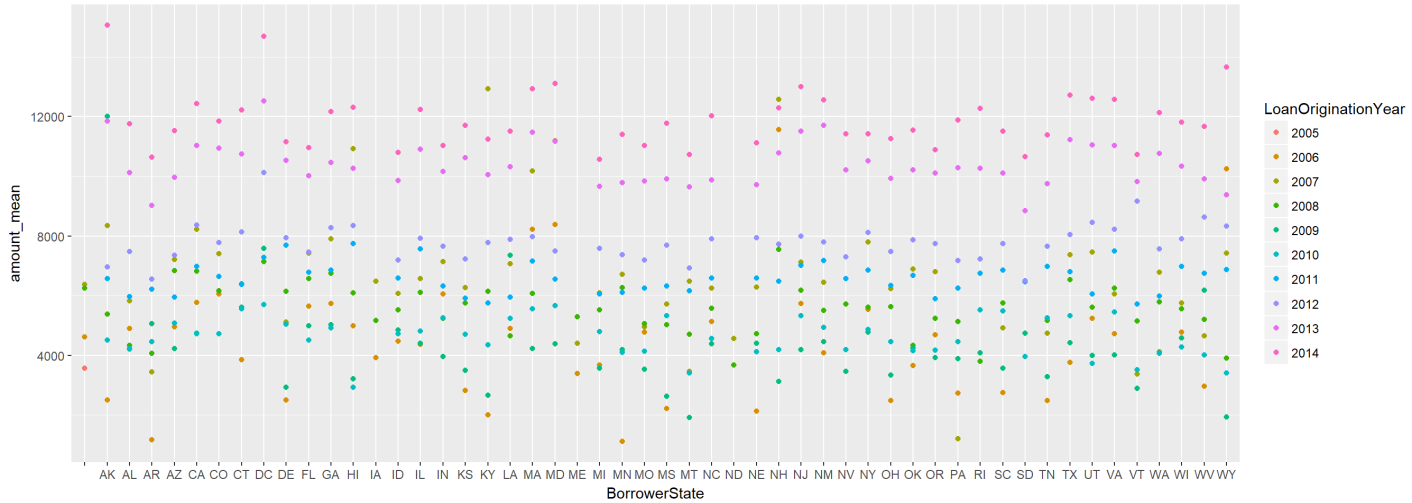
The number of investors followed an increasing trend till 2008 and then decreased for the next three years, which coincides with the period of recession. And then again follows an increasing trend from 2011 onwards.

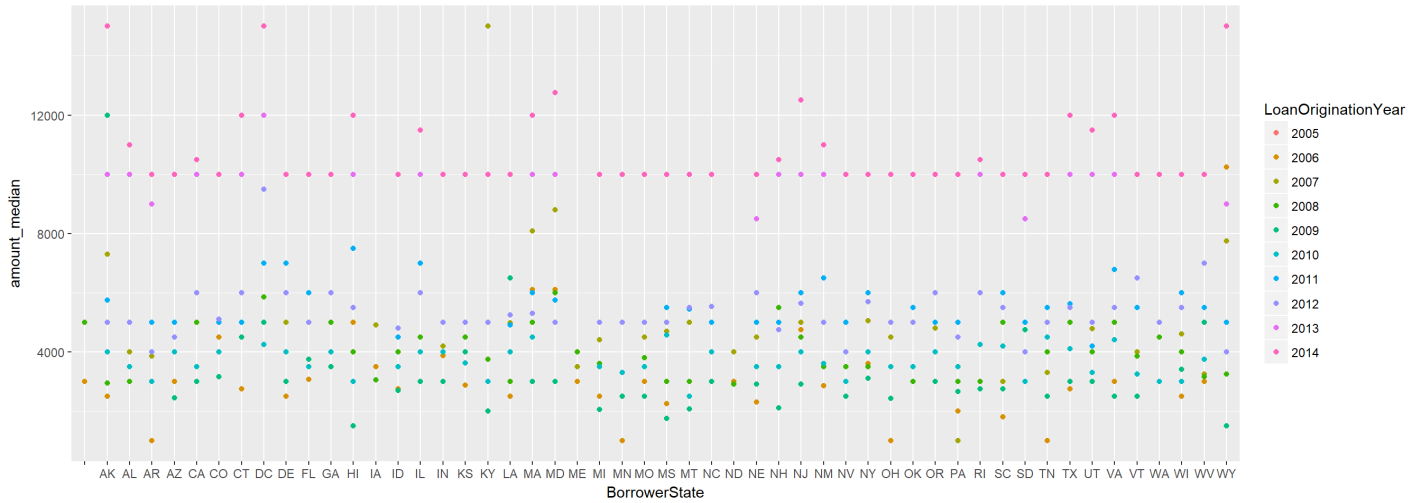
There is an unusual drop in median Loan Original Amount for the year 2012.

The plot between Loan Amount and ProsperScore for median value looks interesting but Pearon correlation value here is very small.

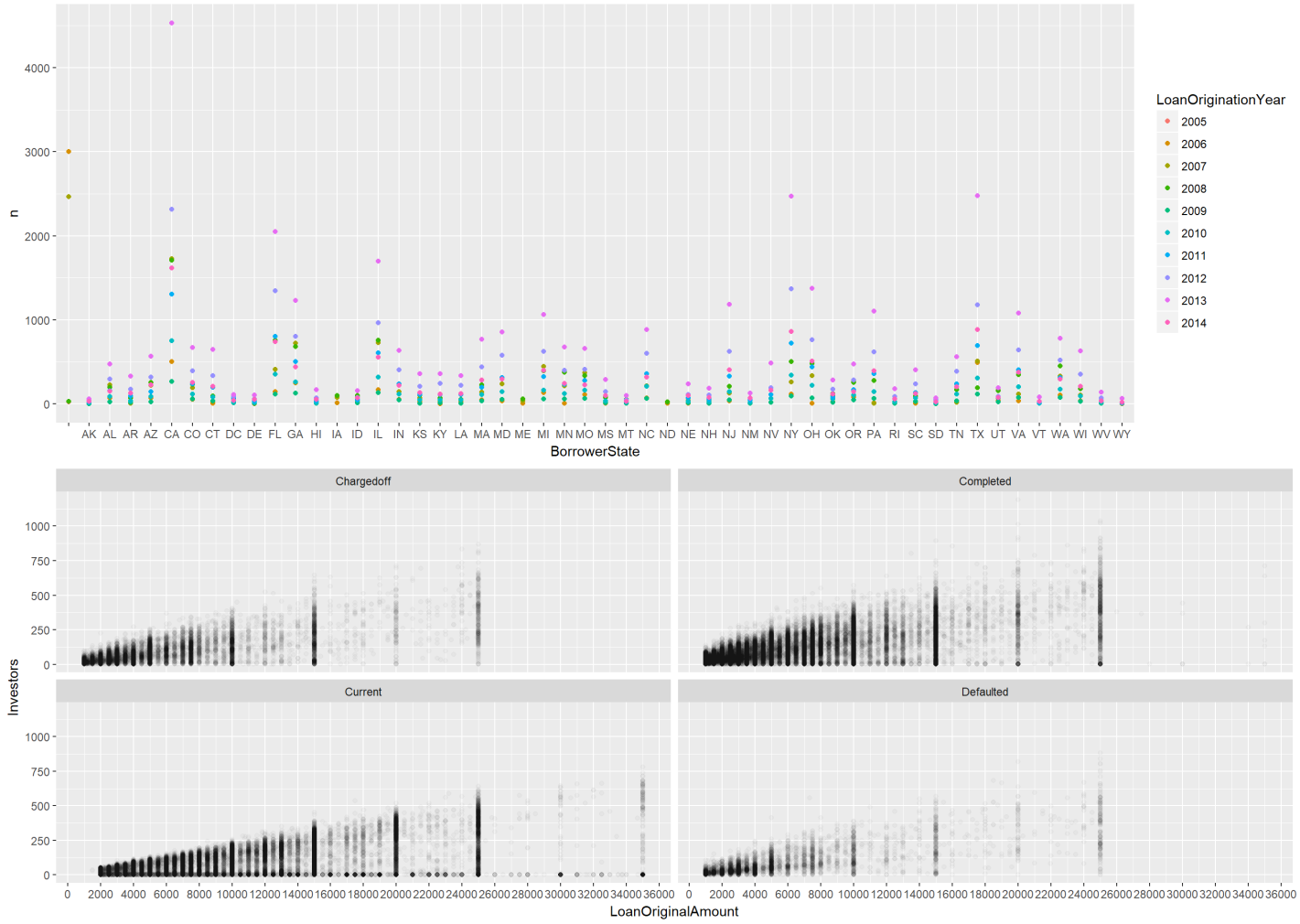
Multivariate Plots Section

```
## # A tibble: 6 x 5
##   BorrowerState LoanOriginationYear amount_mean amount_median     n
##   <fct>         <chr>                     <dbl>         <dbl> <int>
## 1 ""            2005                      3577.         3000.    22
## 2 ""            2006                      4619.         3001.  3000
## 3 ""            2007                      6384.         5000.  2462
## 4 ""            2008                      6256.         5000.    31
## 5 AK            2006                      2500.         2500.     2
## 6 AK            2007                      8345.         7300.    11
```






```
## # A tibble: 10 x 4
##    LoanOriginationYear max_median min_median range_median
##    <chr>                    <dbl>      <dbl>        <dbl>
##  1 2005                     3000.      3000.           0.
##  2 2006                    10250.      1000.        9250.
##  3 2007                    15001.      1000.       14001.
##  4 2008                     6001.      2900.        3101.
##  5 2009                    12000.      1500.       10500.
##  6 2010                     5000.      2500.        2500.
##  7 2011                     7500.      4200.        3300.
##  8 2012                     9500.      4000.        5500.
##  9 2013                    12000.      8500.        3500.
## 10 2014                    15000.     10000.        5000.
```




```
##
##  Pearson's product-moment correlation
##
## data:  prosper$LoanOriginalAmount and prosper$Investors
## t = 138.71, df = 113940, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3751140 0.3850494
## sample estimates:
##       cor
## 0.3800926
```

MultiVariate Analysis:

Plot 1 features: The mean of Loan Original Amount vs Borrower state plot gives the big picture of all the lisitings' loan amounts variation for different states in a single graph.The plots for each year show that the mean loan amount has been increasing with each year. But the points for the years 2005-2011 seem to be mixed up. The top 3 years data is nice and clear.Lots of high points can be seen for the years 2007 and 2006 for unexpected states like NH,WY,MD and HI .Lowest points in the plot also belong to these years for the states AR and MN and PA.

Plot2 features: The median amount has been moving up the y-axis with each year. The first part that catches our eye in this plot is the points at $10,000 mark on y-axis. Median amounts of two years 2013 and 2014 coincide for some states. Again some unusual peaks can be observed in this plot too for states like AK,DC, HI, KY, NH and

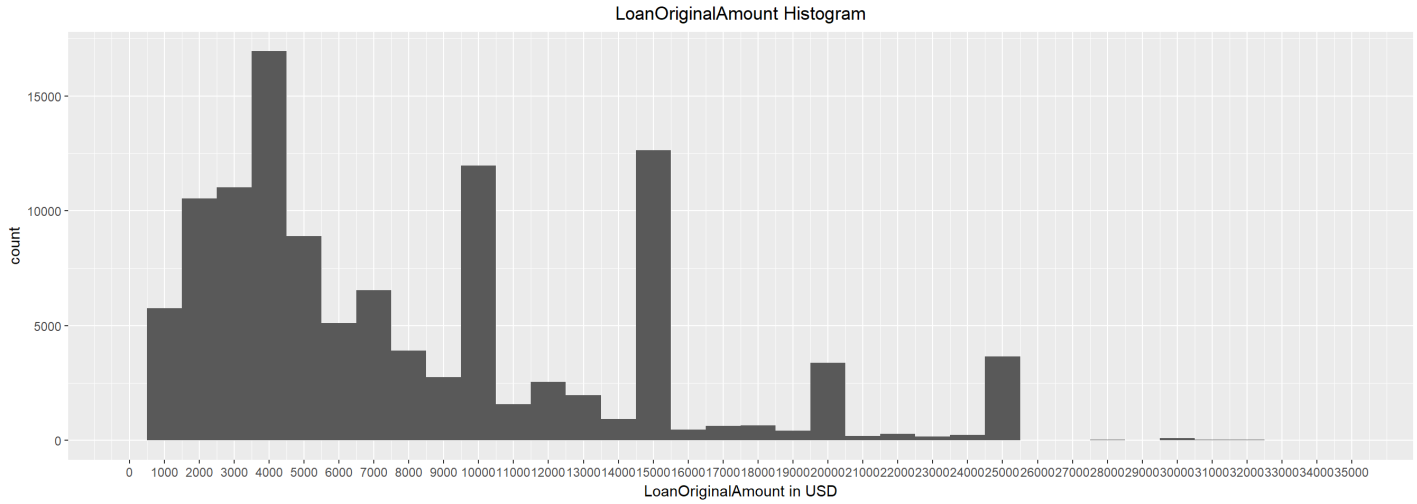
WY. Most of these peaks belong to the years 2006 and 2007. Interestingly, the lowest points in the plot also belong to these two years for states like AR,MN,OH, PA and TN.

Plot3 features: The third plot shows the variation of number of borrowers across different states for different years.Over the given time period,the number has been on an increasing trend and growth was remarkable in the past 5 years for states CA, FL, IL, NY and TX. Whereas for the inital few years, the points can be seen mixed up and have small borrower numbers.

Plot4 features: Clearly defined lines can be seen at regular numbers and the number of investors is nearly uniformly increasing withn the Loan Amount. Also, a huge proportion of listings have their loan amounts under $10,000 and the density decreases with the Loan Amount.Of all the loan status, current loans have the highest range for Loan Amount. Pearson correlation test shows that there is a small correlation between these two variables.

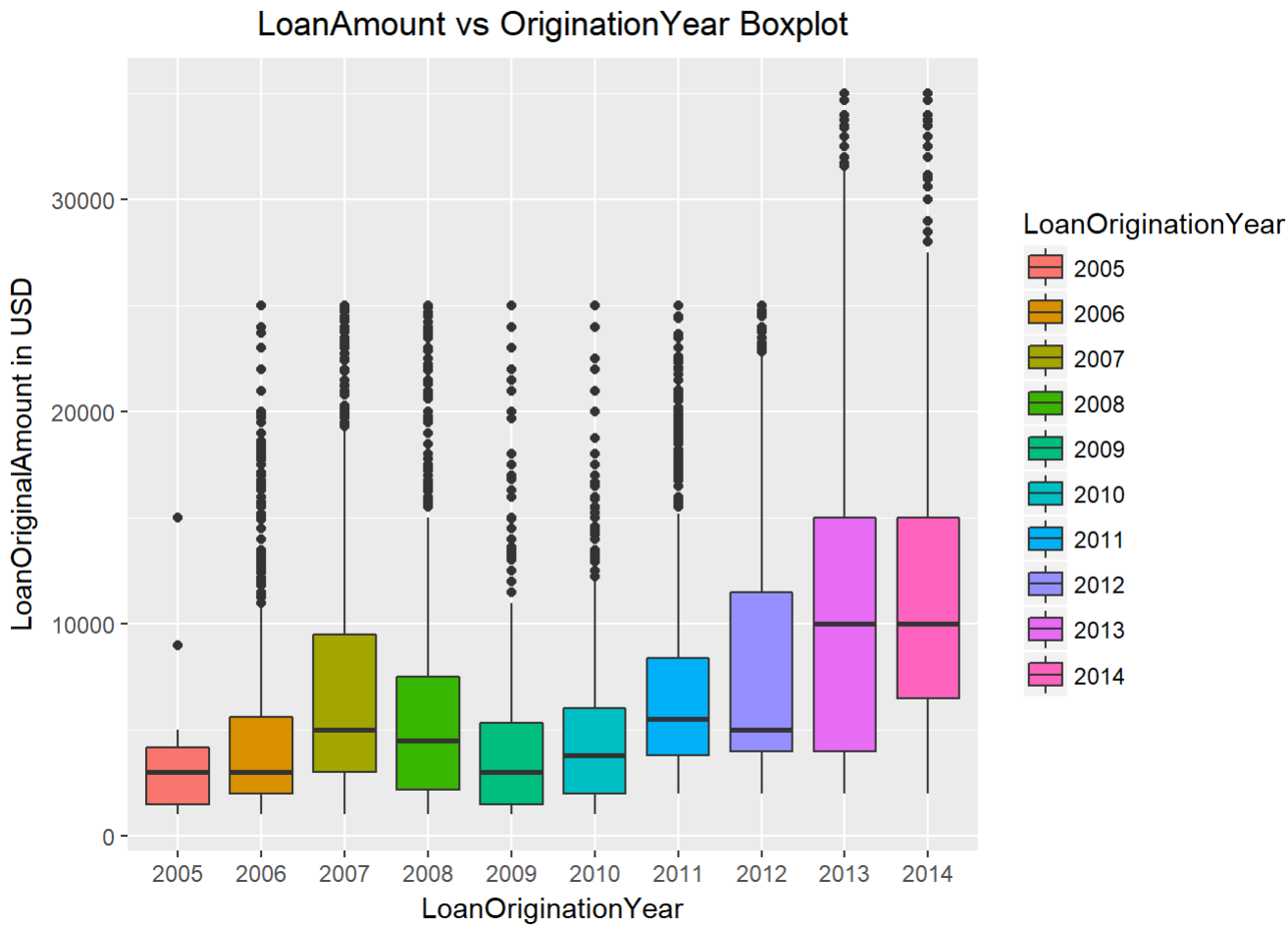
Final Plots and Summary

Plot One:



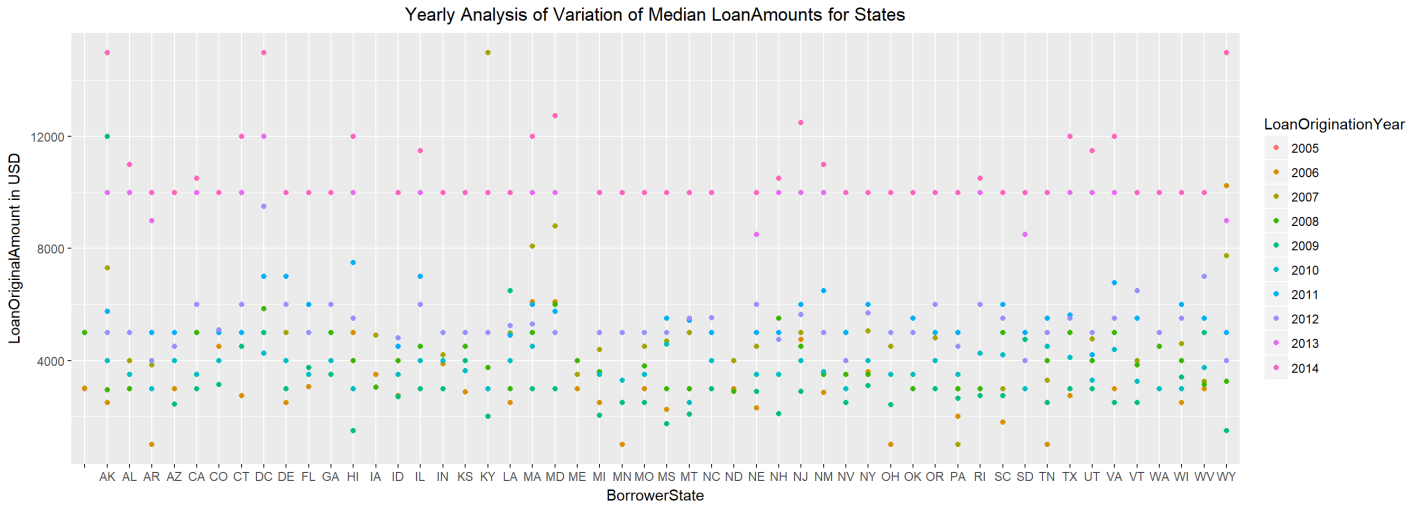
Description One: $5000, $10,000 and $15,000 are the highest frequent amounts borrowed with an overall median amount of $6500.

Plot Two:



Description Two: Median Loan amount touched $5000 for yeara 2007 and then decreased to a minimum value in 2009. From there on, there is a normal increase except for the drop in year 2012 the reason for which remains unknown.Data for year 2014 is available only for 3 months.

Plot Three:



Description Three: The median amoount line has been moving up the y-axis with each year.Some unusual peaks can be observed in this plot for states like AK,DC, HI, KY, NH and WY. Most of these peaks belong to the years 2006 and 2007. Interestingly, the lowest points in the plot also belong to these two years for states like AR,MN, OH, PA and TN.

Reflection

Prosper Loans Dataset contains loan info and borrowers info of nearly 114,000 listings with 10 variables from around November 2005 to March 2014. I started by understanding the individual variables in the data set, and then I explored interesting questions as I continued to make observations on plots.Eventually, I explored the dependencies of different variables of borrower data and analysed the yearly progress made by Prosper.

There is noticeable correlation between Number of Investors and Loan Amounts. Smaller loan amounts tend to attract more number of investors than larger ones.Remarkable growth in business can be seen for states like NY, CA, TX and FL. Overall growth for all states has seen a jump in the year 2013.$5000, $10,000 and $15,000 are the highest frequent amounts borrowed with an overall median amount of $6500.

Some limitations of this analysis include insufficient data for year 2014.It would have been more apt if data was available for atleast three more years and the current year 2018. Sufficient data leads to improved and more useful analysis. Reducing the number of values for categorical variable ‘LoanStatus’ has made the exploration smooth with reduced complexity. As this is a completely new domain to me, I stuck with the basic and easily interpretable variables amongst the many available. To further add to this analysis in future, I would like to analyse data related to defaulted, charged off and cancelled loans to predict the risk factor involved and probe into minimising for these type of listings.